Recognizing emotions in dialogues with acoustic and lexical features

Recognizing Emotions in Dialogues with Acoustic and Lexical Features

Leimin Tian
School of Informatics
the University of Edinburgh
Edinburgh, UK, EH8 9AB
Email: s1219694@sms.ed.ac.uk

Johanna D. Moore
School of Informatics
the University of Edinburgh
Edinburgh, UK, EH8 9AB
Email: j.moore@ed.ac.uk

Catherine Lai
School of Informatics
the University of Edinburgh
Edinburgh, UK, EH8 9AB
Email: clai@inf.ed.ac.uk

***Abstract*—Automatic emotion recognition has long been a focus of Affective Computing. We aim at improving the performance of state-of-the-art emotion recognition in dialogues using novel knowledge-inspired features and modality fusion strategies. We propose features based on disfluencies and non-verbal vocalisations (DIS-NVs), and show that they are highly predictive for recognizing emotions in spontaneous dialogues. We also propose the hierarchical fusion strategy as an alternative to current feature-level and decision-level fusion. This fusion strategy combines features from different modalities at different layers in a hierarchical structure. It is expected to overcome limitations of feature-level and decision-level fusion by including knowledge on modality differences, while preserving information of each modality.**

***Keywords*—emotion recognition; disfluency; dialogue system**

I. Motivation

Research in cognitive science has shown that emotions are vital in human cognition and communication processes [1]. Therefore, it is also important for research in Artificial Intelligence to model emotional intelligence. This led to the establishment of the field of Affective Computing, in which emotion recognition has been a focus. It has become increasingly apparent that automatic recognition of emotion is crucial for advancing technologies related to human-computer interaction, such as human-agent dialogue systems. For example, a virtual agent that is able to copy and adapt its laughter and expressive behaviour has been shown to increase users' humour experience [2]. Similarly, in affective game design, Non-Player Characters that are aware of the emotional states of the player and can generate emotional reactions have been shown to keep players engaged and improve their gaming experience [3]. In a teaching scenario, a robot lecturer expressing a positive mood while giving lectures increased the arousal and positivity of the audience, as well as its perceived lecturing quality [4].

However, the performance of current emotion recognition models is still limited. To address this issue, we proposed novel features based on disfluencies and non-verbal vocalisations (DIS-NVs) in utterances. Our hypotheses are that these features will be predictive for emotion recognition, and that adding them into state-of-the-art models will yield performance improvements.

Features used in emotion recognition can be extracted from various modalities (e.g., audio, visual, and lexical). Our work focuses on features describing the acoustic and lexical characteristics of the dialogues. Previous studies identified Low-Level Descriptor (LLD) based acoustic features and bag-of-words style lexical features as the most predictive for emotion recognition. However, disfluencies and non-verbal vocalisations (DIS-NVs) are also important phenomena in human speech. Evidence from psycholinguistic studies shows that emotions can influence the neural mechanisms in the brain, and thus influence sensory processing and attention [5]. This in turn influences speech processing and production, which may result in disfluencies [6]. Research has also shown that DIS-NVs are indicators of the uncertainty of the speaker [6] and level of conflict in dialogues [7], which are behaviours closely related to emotion. Thus, DIS-NVs may be useful cues for emotion recognition. Therefore, we propose features describing occurrences of DIS-NVs in utterances, and study their predictive power compared to state-of-the-art prosodic and lexical features.

Because DIS-NVs provide additional information, we have good reason to believe that including these features in current emotion recognition models using prosodic and lexical features will improve performance. There are currently two main approaches for combining multiple modalities: feature-level and decision-level fusion. However, there are limitations to both of these fusion strategies and the performance improvement is often moderate [8].

In feature-level fusion, features from different modalities are first concatenated, and then an emotion recognition model is built with this concatenated feature set. A limitation of feature-level fusion is that it does not distinguish features from individual modalities. Thus, incorporating knowledge of differences between individual modalities is difficult in feature-level fused models. For example, research has shown that in the audio modality, there are two-way confusions between sadness and dislike. In the visual modality, there are two-way confusions between sadness and surprise, and between anger and dislike, but sadness and dislike are distinct [9]. When building a multimodal model with feature-level fusion, it is difficult to give flexible weights to the audio and visual features. Thus, modality specific information is hard to apply to feature-level fused models. Ideally, with enough training data, the model will be able to automatically learn weights to represent this knowledge. However, emotional databases are often small in size because of the high cost of emotion annotation. Thus, there may not be sufficient data for the model

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

to automatically learn the optimal weights.

In decision-level fusion, unimodal models are first built with each feature set, and the final decision is made based on the predictions given by each unimodal model. Compared to feature-level fusion, it is easier to incorporate knowledge of individual modalities with decision-level fusion. However, detailed information from each unimodal model is not used by a decision-level model. For example, for a model combining the audio, visual, and lexical unimodal models with a majority voting strategy, if both the audio and lexical models predict class 1, while the visual model predicts class 0, the final decision will be class 1. However, the audio and lexical models may have low confidence and predict the wrong class, while the visual model may have high confidence while predicting the right class. In this case, the final decision gives the wrong prediction. This limitation of decision-level fusion can be reduced by applying better strategies, such as using probabilities of unimodal predictions, or using knowledge-inspired rules. However, information about individual features is still not included in the probabilities. Our knowledge of how humans recognize emotions is also limited, and therefore performance of rule-based models motivated by this knowledge will also be limited.

Therefore, we propose a novel hierarchical fusion strategy as an alternative to feature-level or decision-level fusion. This approach combines features from different modalities at different layers of a hierarchical structure. For example, noisy frame-level features can be incorporated at the bottom layer, while abstract utterance-level features such as the DIS-NV features can be included at a higher layer. We hypothesize that this model will be able to effectively make use of knowledge from different modalities, while preserving information from each modality when making predictions. Thus, it is expected to attain better performance than feature-level and decision-level fusion.

With our hierarchical fusion strategy, we hypothesize that fusing our DIS-NV features with state-of-the-art prosodic and lexical features will result in improved performance over current emotion recognition in dialogues. This emotion recognition model also holds the promise to improve user experience with interactive systems, such as human-agent dialogue systems or affective games.

II. Background and Related Work

A. Psycholinguistic Studies

1) Emotion Theories: There are on-going debates in psychology on how to define, study, and explain emotions. Four major approaches have influenced computational studies of emotions: the Darwinian, Jamesian, cognitive, and social constructivist perspectives [10]. The Darwinian perspective argues that emotions are products of evolution, and focuses on identifying a set of primitive and universal emotion categories (e.g., [11]). The Jamesian perspective argues that emotions are caused by physiological and bodily changes, and focuses on identifying the physiological aspects of emotions (e.g., [12]). The cognitive perspective argues that changing of emotional states are induced by events, and focuses on identifying primitive emotional dimensions and building emotional reaction models (e.g., [13]). The social constructivist perspective studies the cultural, gender, and other individual differences in perceiving and expressing emotions (e.g., [14]).

Many current automatic emotion recognition studies follow the Darwinian emotion theory, which defines emotions in terms of several primary and universal categories, such as Ekman's Big-6 emotion categorization [11]. However, our work focuses on the cognitive emotion theory, which associates emotions with specific appraisals (stimuli that evoke changes in emotional states) and use a set of primitive appraisal components or dimensions to define emotions. This is because our goal is to build emotion recognition models that can be applied to the emotional interaction module of human-computer interaction systems, which are mostly developed with appraisal-based emotion models.

In our work, we use four common emotional dimensions that have been identified as able to describe most everyday human emotions [15]: Arousal, Expectancy, Power, and Valence. The Arousal dimension describes the activeness of the subject; the Expectancy dimension describes whether the subject feels that the things under discussion are predictable (positive values) or surprising (negative values); the Power dimension describes whether the subject feels that (s)he dominates the conversation (positive values) or (s)he is being dominated (negative values). The Valence dimension describes whether the subject has positive feelings (positive values) or negative feelings (negative values) towards the topics under discussion. Values on the emotional dimensions can either be continuous real numbers or discrete scores.

2) Human Emotion Recognition: The emotional state of a person during a conversation tends not to change rapidly and thus depends on the context. Humans convey and perceive emotions through all communicative modalities. When recognizing emotions, human subjects are shown to have better performance when given information from multiple modalities [16]. These findings indicate that a contextual and multimodal model may have better emotion recognition performance. In our current work, we focus on emotion recognition from the audio and lexical modalities, which includes acoustic information such as prosodic and spectral features, and contents of the speech.

Psycholinguistic studies have shown that prosodic cues and the lexical content of the speech are important in human emotion recognition (see [16] for a survey). However, current studies on human-human dialogues suggest that disfluency conveys information such as uncertainty [6], which relates to the Expectancy emotional dimension. Non-verbal vocalisations, especially laughter, have also been identified as universal and basic cues in human emotion recognition [17]. Thus, we propose several DIS-NV features to study the predictiveness of DIS-NV for emotion recognition.

B. Automatic Emotion Recognition in Dialogues

There are three important aspects for building an emotion recognition model: the data, the feature set, and the classification or regression model. For the data aspect, there are two main approaches for collecting conversational emotional databases: by recording acted or spontaneous dialogues.

For the feature aspect, there are two main types of features we can extract for most modalities (e.g., acoustic or visual):

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

knowledge-inspired features describing cues that were identified in psychological studies of human emotion recognition, and statistical features describing properties of the data.

For the model aspect, from a temporal view, models may use information from only the current time, or they can include contextual information; From a structural view, models may be flat using the input feature representations directly, or layered, designed to learn a better feature representation before performing classification or regression. Whether to choose one approach or the other, or to combine them, are questions faced by most emotion recognition researchers. In this work, we attempt to provide a better understanding of these issues by using the following as examples to compare these approaches:

- Data:
  - Spontaneous: the Audio/Visual Emotion Challenge 2012 (AVEC2012) [18]
  - Acted: the Interactive Emotional Dyadic MOtion CAPture database (IEMOCAP) [19]
- Features:
  - Knowledge-inspired: disfluencies and non-verbal vocalisations (DIS-NV) [20]
  - Statistical: Low-Level Descriptors (LLD) [21]
- Model:
  - Contextual:
    - Non-contextual:Support Vector Machine (SVM)
    - Contextual: Long Short-Term Memory Recurrent Neural Network (LSTM)
  - Structural:
    - Flat: SVM [22]
    - Hierarchical: LSTM [23]

A detailed study of these factors is important for developing emotion recognition models. However, most studies rely on intrinsic measures to evaluate different approaches (e.g., correlation coefficients or classification accuracies). This leads to another important question in emotion recognition: will performance improvements shown in intrinsic tests of emotion recognition models result in improvements in emotional interaction quality (e.g., higher engagement and satisfaction of the user), when the emotion recognition model is applied to a human-computer interaction system? We plan to work on this question in the future if we have an available dialogue system to apply our emotion recognition models to.

1) Databases: Here we introduce the AVEC2012 database of spontaneous dialogues [18] and the IEMOCAP database of acted dialogues [19] as examples of state-of-the-art emotional databases. They are the most widely used databases of English dialogues annotated with dimensional emotions.

The AVEC2012 database [18] contains the Solid-SAL part of the SEMAINE corpus [24]. It includes approximately 8 hours of audiovisual recordings and manual transcripts of 24 subjects conversing with 4 on-screen characters with specific personalities role-played by human operators. Each dialogue session is approximately 5 minutes long. Emotions in the AVEC2012 database were annotated as real-value vectors in the Arousal-Expectancy-Power-Valence emotional space. Annotations were provided at the word-level and the frame-level.

The IEMOCAP database [19] contains approximately 12 hours of audio-visual recordings from 5 mixed gender pairs of actors. The recordings were manually transcribed. Each conversation was approximately 5 minutes long. There are two types of dialogues in the IEMOCAP database, non-scripted dialogues and scripted dialogues. When collecting the non-scripted dialogues, the actors were instructed to act out emotionally intense scenarios. When collecting the scripted dialogues, the actors would follow pre-written lines. Emotions were annotated at the utterance-level with a 1 to 5 integer score of the Arousal, Power, and Valence emotional dimensions.

2) Features: Previous work on both the AVEC2012 and the IEMOCAP databases have focused on LLD features for the acoustic model (e.g., [25], [26]). However, there are results indicating that knowledge-inspired features, such as global prosodic features, may also be more predictive (e.g., [27], [28]).

3) Models: Most widely used classification or regression algorithms, for example, Support Vector Machines [29], Hidden Markov Models [27], and Conditional Random Fields [30], have been applied to building emotion recognition models. There have also been studies on feature engineering for emotion recognition, such as Canonical Correlation Analysis [31], and Correlation-based Feature-subset Selection [20]. Although many different algorithms exist and it is important to choose the appropriate one for a specific task, previous work has suggested that the predictiveness of features may have greater influence, and there may not be significant differences between the performance of different machine learning algorithms when using the same feature set under similar circumstances [32].

In recent years, deep learning models have obtained leading performance in machine learning tasks, especially in the areas of computer vision and speech recognition [33]. The network structure of deep learning models allows flexible control when fusing multiple modalities and including contextual information, which enables the models to learn better feature representations automatically. They have also achieved improved performance in emotion recognition compared to conventional machine learning algorithms. For example, deep hierarchical neural networks obtained the best reported results in detecting the Valence emotional dimension values and level of conflict [34], and the use of autoencoders has improved unsupervised domain adaptation in affective speech analysis [35].

However, compared to databases used for speech or image recognition tasks, the emotional databases are relatively small. This may limit optimization of the complex model structure of a deep learning model. The ability to generalize over different databases is also an issue for current deep learning models. In this work, we use the LSTM-RNN model as an example to investigate the predictiveness and robustness of deep learning models for emotion recognition, and compare their performance with the widely used SVM model.

Most recognition models on the AVEC2012 database use Support Vector Regression without including contextual information. However, models that included contextual information, in either the features extracted [20] or the recognition model used (e.g., Hidden Markov Model [27], and Particle Filtering [25]), have shown better performance in emotion recognition. To the best of our knowledge, the only previous work on the AVEC2012 database that applies deep learning models used

the LSTM model to learn better feature representations, and then applied Support Vector Regression on the outputs of the LSTM models [36]. Their results show that the LSTM model learned better feature representations. LSTM models were used directly for classification in previous work on the IEMOCAP database, and they obtained better performance than Hidden Markov Models (e.g., [26], [37]). Another application of deep learning methods uses Denoising Autoencoders to model gender information, which is shown to help with the emotion recognition task [38].

Because different settings were used in previous work, such as data preprocessing and focusing on different emotion annotations, it is hard to compare results. The different nature of emotion recognition tasks on the AVEC2012 and the IEMOCAP database also means that results on these two databases are not directly comparable. Thus, we build our own models using different features and classification methods under the same experimental settings for comparison.

III. Methodology

A. Features

We study three types of acoustic and lexical features in current work. The DIS-NV features and the PMI lexical features are knowledge-inspired features, which describe data at the utterance-level with a small feature set. The LLD features are statistical features, which describe data at the frame-level with a large feature set. The statistical features are able to give detailed information of all the data, while the knowledge-inspired features can highlight the utterances that may be specifically interesting for emotion recognition. The statistical features are data oriented, while the knowledge-inspired features describe general emotional cues. Thus, the knowledge-inspired features may generalize better to unseen data than the statistical features.

1) DIS-NV Features: We study three types of disfluencies: filled pauses (non-verbal insertions, e.g., "eh"), fillers (verbal insertions, e.g., "you know"), and stutters (involuntarily repeats of part of a word or words); as well as two types of non-verbal vocalisations: laughter and audible breath. We manually annotated DIS-NVs for both databases. Feature values are calculated as the ratio between the sum duration of each type of DIS-NV appearing in an utterance and the total duration of the utterance [20]. Filled pauses and laughters were the most predictive types of DIS-NV in our previous studies. We also tested other types of common DIS-NVs, including speech repairs, silent pauses, turn-taking times, sighs, and prolongations. However, adding them to the DIS-NV feature set does not improve recognition performance, thus are omitted. We plan to further study the differences between different DIS-NVs using descriptive statistics.

Compared to the AVEC2012 database of spontaneous dialogues, DIS-NVs are less frequent in the IEMOCAP database of acted dialogues. This indicates fundamental differences between spontaneous and acted dialogues. Frequencies of DIS-NVs in both databases are shown in Table I. In the first row, "FP" is filled pause, "FL" is filler, "ST" is stutter, "LA" is laughter, "BR" is breath. Frequencies of most types of DIS-NV are much lower on the IEMOCAP database. The fillers are the only exception, which may be because some fillers were part of the scripts. Because each pair of actors played out every script, fillers were duplicated when collecting the scripted dialogues of the IEMOCAP database.

TABLE I. Frequencies of DIS-NVs.

Databases	FP(%)	FL(%)	ST(%)	LA(%)	BR(%)
AVEC2012	32.0	*14.7*	9.4	11.9	2.7
IEMOCAP	11.2	*24.1*	6.3	1.6	0.6

2) PMI Lexical Features: The lexical features we extracted are based on Point-wise Mutual Information (PMI). PMI is a widely used measurement for the relation of words and emotions. It is based on the frequency of a word w having class label c: $PMI(c,w) = log_2(\frac{P(c|w)}{P(c)})$. Lexical features based on PMI values were the second most predictive features in previous work on the AVEC2012 database ([20], [25]). Emotional dimensions are binarized when calculating PMI values. The lexical features we proposed are calculated as the total PMI values of all the words in an utterance for each binarized emotional dimension.

3) LLD Acoustic Features: We extracted the LLD acoustic features by using a frame-level sliding window. Functionals (e.g., mean) were applied to LLDs (e.g., MFCCs) and their corresponding delta coefficients. The OpenSMILE toolbox [21] was used to extract these features from audio recordings automatically. The LLD features used for different databases were not exactly the same. As we mentioned in Section II-B3, results on the AVEC2012 and the IEMOCAP databases are not directly comparable, and it is difficult to compare with previous work because of differences in experimental settings. Thus, in our current work we chose the most widely used LLD feature set from previous work on each database as the reference set for experiments on this database. In the future, we will use the union of these feature sets.

B. Classification Models

We build two types of classification models: a SVM model, which does not model sequence information and uses the given feature representations directly; and a LSTM model, which can automatically learn a flexible history length and a potentially better feature representation. Compared to the SVM model, the LSTM model has more parameters that need to be learned during training. Thus, the predictive power of the LSTM model may be limited by the size of the training data. The complex structure of the LSTM model may also risk over-fitting.

1) SVM: Our SVM models were built with the LibSVM [39] classifier using WEKA [40]. We used the C-SVC approach with RBF kernel for both databases. All features were normalized to [-1,1] before classification. This is the setting widely used in previous work (e.g., [18], [29]).

2) LSTM: The LSTM-RNN model is a neural network with multiple hidden layers and a special structure called "the memory cell" that can model long range context information. A hidden layer in a LSTM-RNN model is composed of recurrently connected memory blocks, each of which contains one or more recurrently connected memory cells. Each memory cell has three "gate" units: the input, output, and forget gates. These gates perform the operations of reading, writing, and resetting, respectively. They allow the network to store and retrieve information over long periods of time. The structure

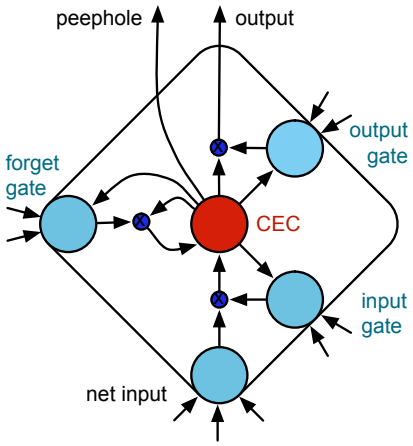

Fig. 1. Structure of a LSTM memory cell [41]

of a LSTM memory cell is shown in Figure 1 [41]. "CEC" in the figure is the "Constant Error Carousel", which is the central neuron that recycles status information from one time step to the next. The small circles with a cross inside are multiplicative connections. The peephole connection gives direct access to the central neuron. Our LSTM-RNN models were built using the PyBrain toolbox [41]. The number of memory cells was selected by cross-validation experiments. All networks were trained using a learning rate of 10^{-5} following settings in [26].

IV. Work Done So Far

Our previous results on the AVEC2012 database verified the predictiveness of the DIS-NV features on spontaneous dialogues. Our experiments have shown that the knowledge-inspired DIS-NV features and PMI lexical features perform better than the statistical LLD features when recognizing emotions in spontaneous dialogues [20]. In contrast, the DIS-NV features and PMI lexical features are less predictive than the LLD features in previous experiments on acted dialogue, which may be because of the infrequency of DIS-NVs in acted dialogues compared to spontaneous dialogues, and the use of scripts in part of the IEMOCAP database [42]. These findings verified that the performance of different types of features vary for different types of dialogues.

Our current results comparing the DIS-NV and LLD features and the SVM and LSTM models on the AVEC2012 and IEMOCAP database are shown in Table II. The numbers are weighted F-measures. "A" is Arousal; "E" is Expectancy; "P" is Power; "V" is Valence; "Mean" is the unweighted average of results on all emotion dimensions. "DN" is using the DIS-NV features. "DN+LLD" represents concatenating the DIS-NV and the LLD features and then applying a LSTM-RNN model to the concatenated feature set.

As we predicted, the complex structure of the hierarchical LSTM model constrains its predictiveness. Because of the small size of the DIS-NV feature set compared to the LLD feature set, the DN-LSTM model has less parameters to train than the LLD-LSTM model. Thus, the DN-LSTM models outperformed the DN-SVM models in both databases, while the LLD-LSTM models are less predictive than the LLD-SVM models in both databases.

Consistent with out previous results, the DIS-NV features are more predictive than the LLD features on the AVEC2012 database of spontaneous dialogues, but not on the IEMOCAP database of acted dialogues. The result that the DN+LLD model only outperformed both the DN-LSTM and the LLD-LSTM model on the IEMOCAP database indicates that a better fusion strategy is needed.

TABLE II. Current results with LSTM models.

Results on the AVEC2012 Test Set						
Features	**Models**	**A** (%)	**E** (%)	**P** (%)	**V** (%)	**Mean**
DN	SVM	52.4	61.4	67.4	59.2	60.1
	LSTM	**54.1**	**65.8**	**68.3**	**60.1**	**62.0**
LLD	SVM	52.4	60.8	67.5	59.2	60.0
	LSTM	52.4	60.7	66.1	58.1	59.3
DN+LLD	LSTM	52.5	61.2	65.8	58.0	59.4
Cross-Validation Results on the IEMOCAP Database						
Features	**Models**	**A** (%)	**E** (%)	**P** (%)	**V** (%)	**Mean**
DN	SVM	36.3	#	40.7	32.8	36.6
	LSTM	41.6	#	37.8	34.0	37.8
LLD	SVM	**65.2**	#	**53.8**	**53.5**	**57.5**
	LSTM	53.7	#	46.2	38.6	46.2
DN+LLD	LSTM	53.9	#	51.6	39.5	48.3

V. Future Plans

In the remaining time of this project, we plan to improve performance of our emotion recognition models by including global prosodic features describing duration, speaking rate, pitch, energy, amplitude, and spectral features of the utterances, as suggested by psycholinguistic studies [43]. We will also extract more robust PMI lexical features based on PMI values calculated from several available corpora with dimensional emotion annotations (e.g., [44]).

We have also conducted pilot experiments on our hierarchical fusion strategy which uses the LLD features at the input layer of the LSTM model and includes the DIS-NV features at a higher hidden layer. Results on the AVEC2012 databases have shown better performance gained compared to feature-level fusion (A=53.4%, E=63.2%). In the next step, we will use the LLD features at the input layer, and the DIS-NV, the PMI lexical, and the global prosodic features at a higher layer in the network structure. We will also compare the performance of hierarchical fusion with decision-level fusion.

If there is an available human-agent dialogue system to apply our models to, in the future we will also perform extrinsic experiments to evaluate the influence of using our emotion recognition models on the quality of interaction.

VI. Contributions

Our work contributes to the Affective Computing community by identifying the predictive DIS-NV features. Our comparison experiments also indicate constraints for selecting among various emotion recognition approaches. The hierarchical fusion strategy we proposed is expected to be a better modality fusion strategy for multimodal models. Applying our model to existing human-computer interaction systems may also improve their quality of interactions.

References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] F. Pecune, M. Mancini, B. Biancardi, G. Varni, Y. Ding, and C. Pelachaud, "Laughing with a virtual agent," 2015.

[3] A. Popescu, J. Broekens, and M. van Someren, "Gamygdala: An emotion engine for games," *Affective Computing, IEEE Transactions on*, vol. 5, no. 1, pp. 32–44, 2014.

[4] J. Xu, J. Broekens, K. Hindriks, and M. A. Neerincx, "Effects of bodily mood expression of a robotic teacher on students," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 2614–2620.

[5] P. Vuilleumier, "How brains beware: neural mechanisms of emotional attention," *Trends in cognitive sciences*, vol. 9, no. 12, pp. 585–594, 2005.

[6] R. Lickley, "Fluency and disfluency," 2015, to appear.

[7] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers." in *INTERSPEECH*, vol. 2005, no. 10, 2005, pp. 1841–1844.

[8] S. D'Mello and J. Kory, "Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 31–38.

[9] L. Chen, H. Tao, T. Huang, T. Miyasato, and R. Nakatsu, "Emotion recognition from audiovisual information," in *Multimedia Signal Processing, 1998 IEEE Second Workshop on*. IEEE, 1998, pp. 83–88.

[10] R. R. Cornelius, "Theoretical approaches to emotion," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[11] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[12] R. W. Levenson, "Autonomic nervous system differences among emotions," *Psychological science*, vol. 3, no. 1, pp. 23–27, 1992.

[13] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[14] E. Spelman, "Anger and insubordination," *Women, knowledge, and reality*, pp. 263–73, 1989.

[15] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[16] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[17] C. McGettigan, E. Walsh, R. Jessop, Z. Agnew, D. Sauter, J. Warren, and S. Scott, "Individual differences in laughter perception reveal roles for mentalizing and sensorimotor systems in the evaluation of emotional authenticity." *Cerebral cortex (New York, NY: 1991)*, vol. 25, no. 1, pp. 246–257, 2015.

[18] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 449–456.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[20] J. Moore, L. Tian, and C. Lai, "Word-level emotion recognition using high-level features," in *Computational Linguistics and Intelligent Text Processing*. Springer, 2014, pp. 17–31.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[22] C. Shawe-Taylor and S. Schölkopf, "The support vector machine," 2000.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1079–1084.

[25] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 485–492.

[26] M. Wollmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of blstm neural networks for enhanced emotion classification in dyadic spoken interactions," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4157–4160.

[27] D. Ozkan, S. Scherer, and L.-P. Morency, "Step-wise emotion recognition using concatenated-HMM," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 477–484.

[28] D. Bone, C. Lee, and S. Narayanan, "Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features," *Affective Computing, IEEE Transactions on*, vol. 5, no. 2, pp. 201–213, 2014.

[29] N. Lubis, S. Sakti, G. Neubig, T. Toda, A. Purwarianti, and S. Nakamura, "Emotion and its triggers in human spoken dialogue: Recognition and analysis," 2014.

[30] T. Baltrusaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[31] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "Cca based feature selection with application to continuous depression recognition from acoustic speech features," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3729–3733.

[32] K. Forbes-Riley and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," *Speech Communication*, vol. 53, no. 9, pp. 1115–1136, 2011.

[33] J. Schmidhuber, "Deep learning in neural networks: An overview," *CoRR*, vol. abs/1404.7828, 2014.

[34] R. Brueckner and B. Schuller, "Be at odds? deep and hierarchical neural networks for classification and regression of conflict in speech," in *Conflict and Multimodal Communication*. Springer, 2015, pp. 403–429.

[35] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," 2014.

[36] J. Wei, E. Pei, D. Jiang, H. Sahli, L. Xie, and Z. Fu, "Multimodal continuous affect recognition based on lstm and multiple kernel learning," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 2014, pp. 1–4.

[37] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling." in *INTERSPEECH*, 2010, pp. 2362–2365.

[38] R. Xia, J. Deng, B. Schuller, and Y. Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 990–994.

[39] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[40] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[41] T. Schaul, J. Bayer, D. Wierstra, Y. Sun, M. Felder, F. Sehnke, T. Rückstieß, and J. Schmidhuber, "PyBrain," *Journal of Machine Learning Research*, 2010.

[42] L. Tian, C. Lai, and J. Moore, "Recognizing emotions in dialogues with disfluencies and non-verbal vocalisations," in *Proceedings of the 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech*, 2015.

[43] I. R. Finlayson, "Testing the roles of disfluency and rate of speech in the coordination of conversation," Ph.D. dissertation, QUEEN MARGARET UNIVERSITY, 2014.

[44] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.